

Figure Legends

Figure 1: Differential expression analysis of all the disease phenotypes in South Africa compared to healthy controls before the initiation of TB treatment. Gene expression profiles of A) TB-only (n=11); B) DM-only (n=33), C) DM-TB (n=15), D) IH-TB (n=20), each relative to healthy controls (n=24). Genes that were deemed statistically significantly differentially expressed had an adjusted $P < 0.05$ after multiple testing correction (Benjamini & Hochberg). Purple corresponds to the genes whose expression was significantly changed, grey shows genes without significant expression change. FDR: false Discovery Rate.

Figure 2: Concordance and discordance of gene expression between the comparisons of each disease group and healthy controls in South Africa. Log fold changes and p-values between groups was calculated with R-package DESeq2. A disco.score was calculated for each pair of corresponding genes. The axes show \log_2 fold change between the conditions indicated by the labels. For example, on the top left plot the X axis corresponds to the comparison between TB and HC, and the Y axis shows the \log_2 fold change between DM-TB and HC. Red dots show genes that are significantly different from the controls in the same direction (concordant genes), and blue dots show genes which are significantly different in both comparisons, but in opposite directions. Intensity of colour indicates the strength of concordance / discordance as measured by the disco.score.

Figure 3: Principal Component Analysis of South African participants. The list of all genes which were significantly differentially expressed in any patient group comparison with healthy controls was used in a Principal Component Analysis of all the samples obtained from participants recruited in South Africa.

Figure 4: Transcriptional modules that were significantly differentially expressed in TB-only, DM-TB, IH-TB and DM-only compared to healthy controls in South Africa before initiation of TB treatment. Transcripts were evaluated using a pre-existing modular framework. Significantly up-regulated (red) and down-regulated (blue) modules are shown: the length of each bar corresponds to the effect size (magnitude of change) of that module, and the colour saturation represents the adjusted p-value (< 0.0001). The amount of colour represents the proportion of genes within that module that were differentially expressed.

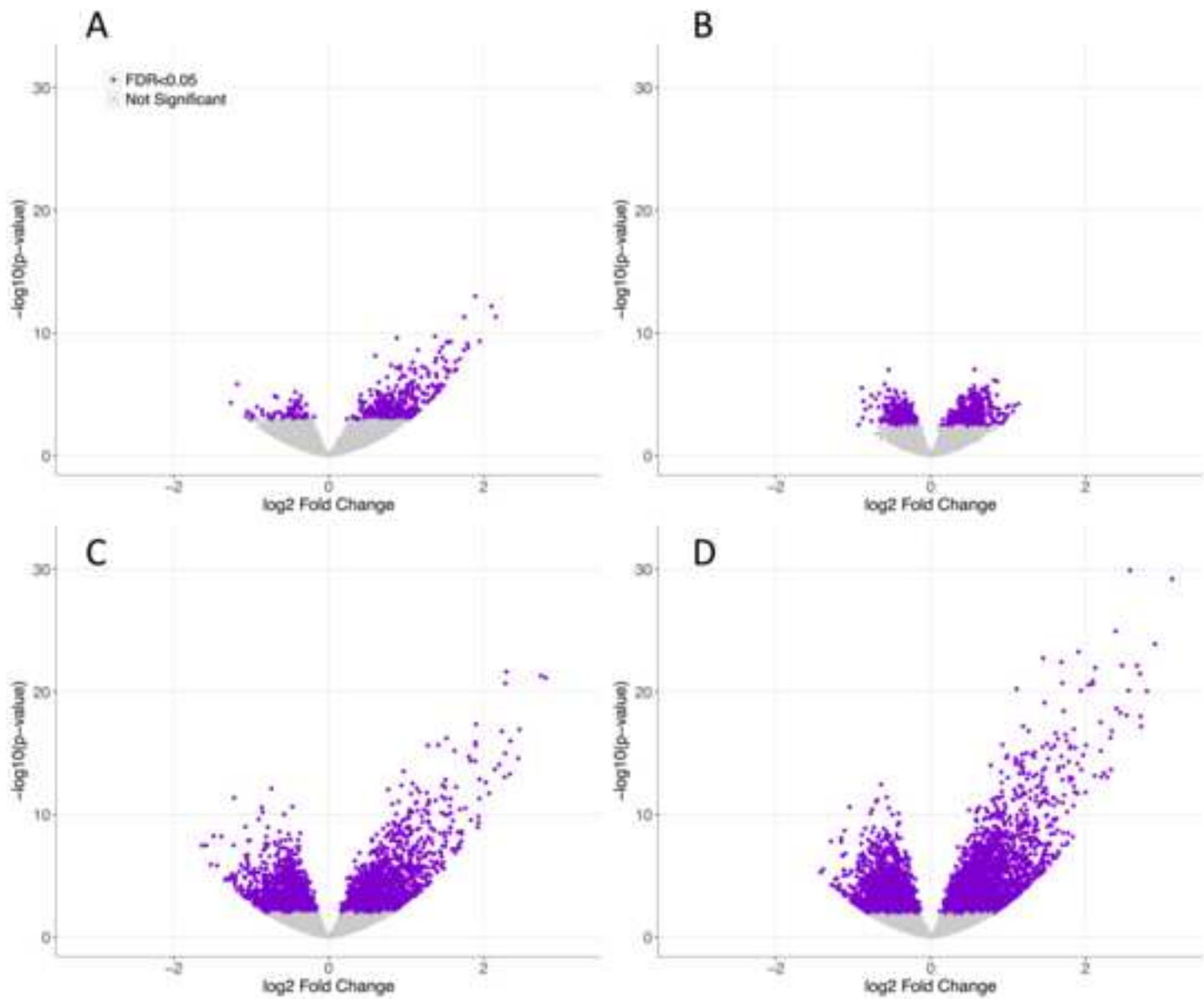
Figure 5: Differential gene expression analysis of DM-TB and IH-TB patients relative to TB-only patients, in the combined dataset from all four field sites. Samples collected in South Africa, Romania, Peru and Indonesia from A) DM-TB patients and B) IH- patients were compared with patients with TB-only in a combined analysis. Genes significantly differentially expressed after multiple testing correction are shown in pink (p-value < 0.05). Genes in grey are not statistically significantly altered compared to patients with untreated TB only. FDR: false discovery rate.

Figure 6: Summary of modular analysis in all four field sites. The fold changes of the genes within the top significantly differentially expressed modules are shown (adjusted p-value < 0.05). On the inside: IH-TB compared to TB-only. Outside: DM-TB compared to TB-only. Up-regulated genes are shown in red, and down-regulated genes are in blue. The saturation of colour represents the magnitude of differential expression.

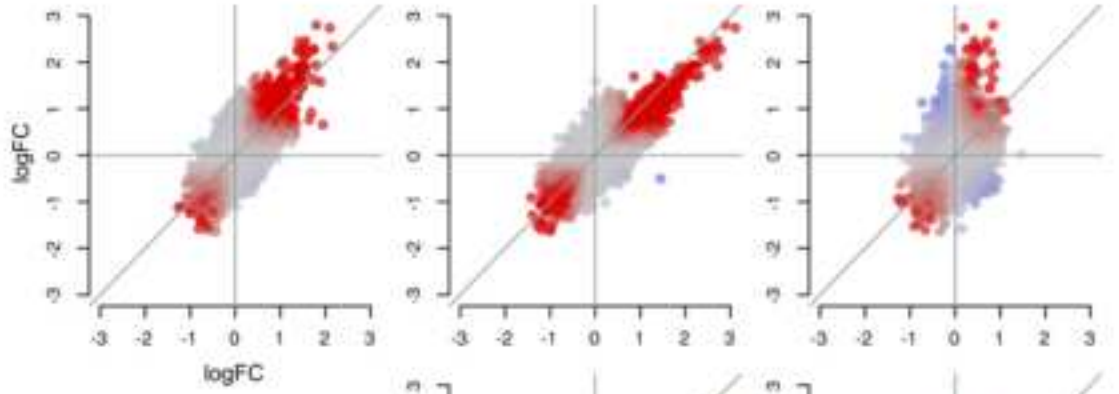
Figure 7: Predictive model of known signature in TANDEM data

Receiver operating characteristic curves based on a machine learning model generated from two different external data set of transcriptome profiles of TB patients and healthy controls (Kaforou [34] and Sweeney [40] training set). The random forest model was applied to the TANDEM cohort (test set), separately to individuals with (“DM”) and without DM (“no DM”).

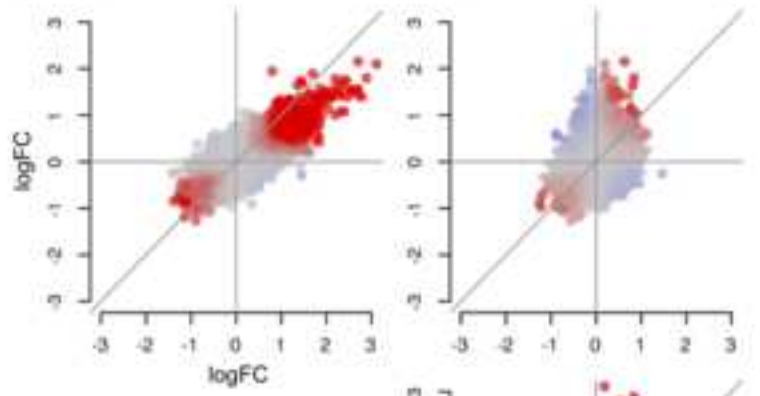
Accepted Manuscript



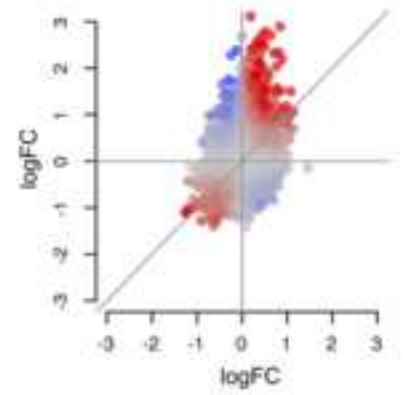
DMTB_v_HC



TB_v_HC



IHTB_v_HC



DM_v_HC

